

HumRRO Report
FR-01-41

August, 2001

Third-Party Checking of 2001 Scaling and Linking for the Kentucky Core Content Test

R. Gene Hoffman
Arthur A. Thacker
Laura A. Ford

Prepared for:

The Kentucky Department of Education
500 Mero Street
Frankfort, KY 40601

Contract Number M-00003669

Third-Party Checking of 2001 Scaling and Linking for the Kentucky Core Content Test

Table of Contents

Introduction.....	1
Scaling and Linking Procedures.....	1
Scope of Third-Party Checking.....	2
Processing Steps.....	2
Results	3
Anomalies.....	5
Documentation.....	5
Conclusion.....	7
References	8

Summary

CTB and HumRRO independently calculated the scaled/linked raw-score-to-scale-score tables for the 2001 Kentucky Core Content Test. From those tables, both identified cutpoints that could be used for assigning student performance classifications and later converted to school accountability indexes. Decisions regarding the handling of problem test items were discussed between CTB and HumRRO and in all cases both groups reached consensus. All results calculated by HumRRO were identical to those calculated by CTB. Given that our scaling and linking results are identical with those of CTB, we can be assured that CTB did not commit processing errors.

Third-Party Checking of 2001 Scaling and Linking for the Kentucky Core Content Test

Introduction

In order to make the transition from the Kentucky Instructional Results Information System (KIRIS) test to the Kentucky Core Content Test with the minimum amount of disruption, a system of linking the old test with the new was necessarily devised. This link allowed Kentucky to maintain consistency in its student performance levels and to apply the student Kentucky Core Content Test scores to a newly revised accountability calculation. The main difficulty in linking the two tests was that KIRIS only applied student scores on the open-response section of the test toward a school's accountability index and toward individual student performance levels. The Kentucky Core Content Test uses both open-response and multiple-choice format questions to make those determinations. Students still receive ratings in terms of the Novice, Apprentice, Proficient, and Distinguished levels of performance, but multiple-choice questions are now included in those determinations. A two-step process was used to make the link from the Kentucky Core Content Test back to the KIRIS scale on which student performance standards had been set in 1993 (Kentucky Department of Education (KDE), 1997). The first step involved analysis of 1998 data in which multiple-choice and open-response items were combined on a single scale and that combined scale equated to the open-response-only scale. HumRRO, in an earlier report (Hoffman, Thacker, & McBride, 1999), performed a third-party evaluation of those procedures. The second step, for linking the Kentucky Core Content Test back to the KIRIS scale, was to link the 1999 test data to the newly created combined scale. HumRRO also performed a third-party evaluation of those procedures (Hoffman & Thacker, 1999).

The 2001 administration of the Kentucky Core Content Test, a major component of the Commonwealth Accountability Testing System (CATS), was also linked back to the combined scale. This was accomplished by linking the 2001 test back to the 2000 test. The procedures for doing so mimic the procedures used in 1999 and 2000 (Hoffman & Thacker, 2000). This report represents HumRRO's third-party check of the scaling and linking of the 2001 Kentucky Core Content Test.

Scaling and Linking Procedures

Item data from all forms were scaled using CTB's PARDUX program. Item parameters were then divided by form and entered into CTB's FLUX program to create raw-score-to-scale-score conversion tables. The scaling process included adjusting item parameters by PARDUX application of the Stocking-Lord procedure to items linking the 2001 Kentucky Core Content Test to the 2000 administration of the test. One form from each grade/subject was identified from the 2000 Kentucky Core Content Test to serve as an anchor form. Each anchor form was readministered in 2001 with all items intact and occurring in the same sequence as in 2000. All anchor item parameters come from the multiple-choice items included on the anchor form. Open-response items were repeated on the anchor forms for form construction consistency and to ensure that contextual clues that may have been present in 2000 were repeated for 2001.

Cutpoints were established over the past year by Kentucky teachers during the most comprehensive standards-setting process ever undertaken by any state, national or international

testing system (<http://www.kde.state.ky.us/comm/pubinfo/standards/>). The teachers used a combination of the Contrasting Groups, Jaeger-Mills, and CTB Bookmark procedures to set standards. The process was designed and overseen by the National Technical Advisory Panel on Assessment and Accountability, which has guided the development of CATS.

Scope of Third-Party Checking

HumRRO conducted parallel analyses to accomplish scaling and linking for the 2001 data.

Processing Steps

HumRRO took the following steps for each grade/subject tested:

1. Verify anchor files (PARDUX *.anc) of multiple-choice test items that appear on the anchor form. These anchor items are used to link the 2001 test to the 2000 scale which was previously adjusted to the 1993 scale. 2001 anchor files were checked against 2000 parameter files for the matching forms.
2. Create working files (PARDUX *.RWO) from the 2001 Kentucky Core Content Test calibration sample. These files include both open-response and multiple-choice data.
3. Prepare control files (PARDUX *.ctl) which contain the constraints used for item parameter estimation, student proficiency estimation, maximum number of items, etc. The SAS program used to create *.rwo files included a routine to print out a control file.
4. Estimate parameters for Kentucky Core Content Test items using PARDUX.
5. Perform Stocking-Lord transformation using PARDUX. The results of this transformation include a slope and intercept constant for linking the 2001 Kentucky Core Content Test back to 2000.
6. Confirm that the equating constants from Step 5 match those derived by CTB.
7. Create parameter files (FLUX *.par) for each test form for use in preparation of raw-score-to-scale-score tables. A special SAS program was written for this purpose.
8. Create files (FLUX *.hbk) containing the scale limits (325 and 800) and constants from the Stocking-Lord transformation. This was a simple word processing task.
9. Create raw-score-to-scale-score transformation tables for each form using FLUX.
10. Confirm that the raw-score-to-scale-score transformation tables from Step 9 match those derived by CTB.
11. Confirm that the cutpoints set by CTB were consistent with established cutpoints from the KDE (<http://www.kde.state.ky.us/comm/pubinfo/standards/>).

Results

After performing periodic checks with CTB as individual tests were scaled and equated, HumRRO and CTB reached exact agreement on the equating constants for all grade/subjects. Table 1 summarizes the results of this study. It identifies the grade and subject for each test in the first two columns. The third column identifies problem items and references the solutions that were reached by CTB and verified by HumRRO. The next four columns contain the M1 and M2 (slope and intercept) constants obtained from the Stocking-Lord transformation. HumRRO computed the first set of constants, CTB the second. The seventh and eighth columns contain the difference between the M1 and M2 constants computed by HumRRO and those computed by CTB.

The last column in Table 1 is a verification of the exact agreement between CTB and HumRRO for the raw-score-to-scale-score tables. Cutpoints from those tables are used to assign students to performance categories (Novice Apprentice, Proficient, or Distinguished), that are in turn used in the computation of each school's accountability index. CTB and HumRRO were in exact agreement for all raw-score-to-scale-score tables for every grade/subject.

The asterisks from the third column of Table 1 represent problem items. Each asterisk is referenced with the specific problem that occurred and the solution. All problem items were dealt with during the parameter estimation phase of the scaling and equating process. No item for which parameters were estimated was eliminated from the Stocking-Lord procedure. The same column indicates whether or not convergence was reached during parameter estimation. If convergence was not reached after 50 iterations by the Pardux program, the solution at stage 50 was accepted by mutual agreement.

HumRRO also verified the cutpoints on the raw-score-to-scale-score tables. Cutpoints were assigned by rule. HumRRO verified cutpoints between Novice and Apprentice, between Apprentice and Proficient, and between Proficient and Distinguished performance categories. HumRRO also verified cutpoints for Low, Medium, and High subcategories within the Novice and Apprentice categories.

Third Party Checking 2001

				CTB		HUMRRO		CTB-HUMRRO Differences			
Grade	Subject	Convergence	Problems	M1	M2	M1	M2	M1	M2	Scor. Tables Exact Agreement	Cut Points Check
4	RD	No ¹	Convergence	30.93623	548.13098	30.93623	548.13098	0.00000	0.00000	YES	YES
	SC	Stage 18	Item 131 ²	25.78811	546.84253	25.78811	546.84253	0.00000	0.00000	YES	YES
5	A&H	Stage 26	None	44.12090	510.82654	44.12090	510.82654	0.00000	0.00000	YES	YES
	MA	Stage 23	None	34.29762	560.51556	34.29762	560.51556	0.00000	0.00000	YES	YES
	PL	Stage 14	Item 40 ³	45.55397	504.00894	45.55397	504.00894	0.00000	0.00000	YES	YES
	SS	Stage 16	None	31.28720	539.07770	31.28720	539.07770	0.00000	0.00000	YES	YES
7	RD	Stage 9	None	29.68446	513.80884	29.68446	513.80884	0.00000	0.00000	YES	YES
	SC	Stage 18	None	25.37041	502.77740	25.37041	502.77740	0.00000	0.00000	YES	YES
8	A&H	Stage 18	None	49.65580	512.27838	49.65580	512.27838	0.00000	0.00000	YES	YES
	MA	Stage 25	None	32.97469	533.97778	32.97469	533.97778	0.00000	0.00000	YES	YES
	PL	Stage 15	None	43.07016	503.29736	43.07016	503.29736	0.00000	0.00000	YES	YES
	SS	Stage 17	Items 141 and 156 ⁴	40.58748	512.67188	40.58748	512.67188	0.00000	0.00000	YES	YES
10	PL	Stage 15	None	45.14058	504.03198	45.14058	504.03198	0.00000	0.00000	YES	YES
	RD	Stage 15	None	51.50084	508.14221	51.50084	508.14221	0.00000	0.00000	YES	YES
11	A&H	Stage 19	Item 89 ⁵	49.39001	516.80975	49.39001	516.80975	0.00000	0.00000	YES	YES
	MA	Stage 30	None	40.25501	535.32367	40.25501	535.32367	0.00000	0.00000	YES	YES
	SC	Stage 23	Item 67 ⁶	30.73619	544.08887	30.73619	544.08887	0.00000	0.00000	YES	YES
	SS	Stage 18	None	48.32495	545.51117	48.32495	545.51117	0.00000	0.00000	YES	YES

¹Convergence was not reached for RD07. The solution at stage 50 was used operationally.
²Parameters could not be estimated for item 131 in SC04. The item was omitted.
³Item 40 in PL05 required an M-step for parameter estimation.
⁴Items 141 and 156 in SS08 both required an M-step for parameter estimation.
⁵Item 89 in A&H11 required an M-step for parameter estimation.
⁶Parameters could not be estimated for item 67 in SC11. The item was omitted.

Anomalies

In one instance HumRRO and CTB calculated differing initial M1 and M2 (slope and intercept) constants. Upon investigation it was determined that 15 students each received blank scores for one open-response item on the fifth-grade Practical Living/Vocational Studies test in CTB's student data file. When HumRRO read the student data, those scores were automatically changed to 0's. They were left as "missing" in CTB's file, causing the production of two different incorrect solutions. By rule, no student should have missing data for an open-response item. Both CTB and HumRRO recalculated their results, omitting the students with missing scores from the calibration sample. Once those students were deleted, CTB's and HumRRO's results matched.

Data Recognition Corporation (DRC) provides student records for both CTB and HumRRO. Once the missing student responses were brought to DRC's attention, they provided a corrected file. When CTB and HumRRO recalculated the results using the newly corrected file, they matched exactly. Results for fifth-grade Practical Living/Vocational Studies were calculated three times by both CTB and HumRRO; first when the results did not match, again omitting the students with missing scores, and finally using a corrected student record file with all student scores included. Table 1 refers to the final analysis.

By rule, blank responses are coded as "B" rather than left blank in the student record file. "B's" then become "0's" during scoring. The exact nature of how the missing data originally occurred is unknown. It is also unknown if missing data also occurs for student records that are not included in the calibration sample. The occurrence of this anomaly is an indication that all student records should be checked for missing data prior to score reporting.

Documentation

To document the steps involved in scaling and linking the 2001 Kentucky Core Content Test we saved all electronic files used in data preparation, including SAS programs, SAS logs, and SAS output lists and all files produced during PARDUX scaling and FLUX transformations. These files have been submitted to KDE. Appendices from the 1999 report (Hoffman & Thacker, 1999) contain printed examples of important files that were submitted.

All electronic files submitted to KDE are named according to the following code (where S = subject, G = grade level).

- A. PARDUX Control File (SSGG01.CTL). This file contains the number of items, the maximum number of stages for PARDUX, the convergence criterion, parameter estimation limits, maximum and minimum values for proficiency estimates (theta), and other information. This file also contains information allowing the program to distinguish between open-response and multiple-choice items, the number of score levels for open-response data, and which items to include in parameter estimation.
- B. PARDUX Data File (SSGG01.RWO). This file contains the student score data. It is coded such that a 1 indicates a correct answer for a multiple-choice question and actual score levels (0-4) are recorded for student responses to open-response questions. To facilitate communication, HumRRO adhered to CTB's item order in constructing these data files.

- C. PARDUX Anchor File (SSGG01.ANC). This file contains 2000 common-scaling item parameters for the 2001 Kentucky Core Content Test. These items were unchanged from 2000 to 2001. Only multiple-choice items are used in *.ANC files.
- D. SAS Programs for Creating Anchor Files, PARDUX Control Files, *.rwo files, and separating parameters by form. The naming convention for these programs is SSGGrwcd.sas and SSGGmakeparfiles.sas. SAS logs and list files are also included from these programs.
- E. PARDUX Parameter Estimation Summary (SSGG01_SUM.TXT). This file provides a summary of the parameter estimation procedure run in PARDUX. It includes the limit data from the control file and also contains the number of stages PARDUX runs in order to reach convergence. It also contains the item numbers of items that could not be estimated and documents any items whose estimation reaches the maximum alpha parameter. This file identifies any problem items that might require additional manipulation before continuing the process.
- F. PARDUX Parameter Estimation Details (SSGG01_DET.TXT). This file is a thorough iteration of the item data during each stage of parameter estimation.
- G. PARDUX Parameter File (SSGG01.PAR). This file contains parameter estimates for all items designated by the *.CTL file. It is used for later data manipulation.
- H. PARDUX Item Summaries Files, Status (SSGG01_STAT.TXT). This file lists all items for a given test and their status after parameter estimation. Items are coded as either estimate OK, OK—default C, not estimated, or other codes. This file provides a different type of record for the parameter estimation.
- I. PARDUX Item Summaries Files, Distribution (SSGG01_DIST.TXT). This file contains the distribution of students who scored at each level on the open-response items. This file is useful for examining the way that scoring rubrics for these items operate and for ensuring that all open-response items have the correct number of functioning score levels.
- J. PARDUX Item Summaries Files, Parameters (SSGG01_PAR.TXT). This file contains the item parameters in a more readily edited format than the *.PAR file. This file can easily be read into word processors and spreadsheet programs.
- K. PARDUX Item Summaries Files, Standard Errors (SSGG01_SE.TXT). This file contains the standard errors of estimation for each item including the errors for the various score levels on the open response items.
- L. PARDUX Item Summaries Files, FitQ1 (SSGG01_Q1.TXT). This file contains fit statistics for all items.
- M. PARDUX Log File (SSGG01_LOG.TXT). As each manipulation of data is completed, PARDUX maintains a log of the procedures and filenames. This log is saved in text format.
- N. Stocking-Lord Plots (SSGG01_SLPLOTS.doc). The Stocking-Lord transformation of the data, which provides the M1 and M2 values (slope and intercept) that allow for the later

creation of scoring tables, outputs three graphs (one each for the a, b, and c parameters) for each transformation. In this file the four graphs (for a, b, c, and P parameters) that result from the transformation using all anchor items are included. The document also contains M1 and M2 transformation constants and a log from the Stocking-Lord procedure.

- O. FLUX control file (SSGG01.HLK). This file specifies the range of the scale scores as well as the M1 and M2 transformation constants to be used from the Stocking-Lord transformation.
- P. FLUX Parameter Files by Form (SSGG01FORM1A.PAR, etc., one for each Form). Each of the parameter files computed using PARDUX was divided to represent items from each test form. Typically, 30 items were scored from each form. Arts and Humanities and Practical Living/Vocational Studies forms contained 10 items to be scored.
- Q. Raw-Score-to-Scale-Score Tables (SSGG01RStoSSTables.doc). A raw-score-to-scale-score table was produced for each form. These tables were saved in text format using FLUX.
- R. Additional files and programs may also be included in the documentation. Those files were constructed either for future purposes or during investigation of results. Student records from which all analyses were conducted were provided by DRC and are included as well.

Conclusion

CTB and HumRRO independently calculated the scaled/linked raw-score-to-scale-score tables for the 2001 Kentucky Core Content Test. From those tables, both identified cutpoints that could be used for assigning student performance classifications and later converted to school accountability indexes. No differences were found between CTB's and HumRRO's parameter estimation, Stocking-Lord transformation constants, raw-score-to-scale-score tables, or application of cutpoints. Given that our scaling and linking results are identical with CTB, we can be assured that CTB did not commit processing errors.

References

Hoffman, R. G. & Thacker, A. A. (2000). *Third-party checking of 2000 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report FR-00-39). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Thacker, A. A. (1999). *Third-party checking of 1999 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report SP-WATSD-99-44). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., Thacker, A. A., & McBride, J. R. (1999). *Documentation of third-party checking of 1998 pre-equating for Kentucky Core Content Test: IRT scaling of multiple choice and open response test items*. (HumRRO Report SP-WATSD-99-39). Alexandria, VA: Human Resources Research Organization.

Kentucky Department of Education (2001). About Kentucky's Performance Standards. Retrieved August 17, 2001, from <http://www.kde.state.ky.us/comm/pubinfo/stadards/>.

Kentucky Department of Education (1997, May). *KIRIS accountability cycle 2 technical manual: based on analysis of data from the 1992-93 through 1995-96 school years*. Frankfort, KY: Author.